САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ – ПРОЦЕССОВ УПРАВЛЕНИЯ

Бакытбек Нурсултан

Выпускная квалификационная работа бакалавра

Анализ тональности русскоязычных комментариев в социальных медиа

Направление 010300

Фундаментальная информатика и информационные технологии

Бакалаврская программа «Программирование и информационные технологии»

Научный руководитель, Малинина М. А.
старший преподаватель

Рецензент, Погожев С.В.
кандидат физ.-мат. наук,
доцент

Санкт-Петербург
2017

Содержание

Аннотация

Выпускная квалификационная работа посвящена вопросу методов автоматического определения тональности русскоязычных текстов, основанных на машинном обучении с учителем. Для исследования были собраны русскоязычные комментарии к видеозаписям на сайте Youtube. Рассмотрены способы векторного представления текстов обучающей и тестовой выборок, а также функции весов. Приводится описание алгоритмов классификации: наивного Байесовского классификатора и метода опорных векторов. Проведены расчеты по эффективности алгоритмов.

Ключевые слова: тональность, тональный словарь, сентимент, сентимент-анализ, анализ мнений, определение тональности, лексическая тональность, обработка текста, анализ комментариев, социальные медиа, Youtube, анализ тональности, классификация по тональности, машинное обучение.

Введение

Сегодня в мире бурно развивается процесс формирования информационного общества. Этому способствует быстрое развитие технологий, появление новых каналов коммуникаций, формирование активной гражданской позиции, а также внедрение новых коммуникационных платформ. Интернет создает принципиально новую среду коммуникации, стирая территориальные границы и расширяя возможности общения. Современный исследователь медиапространства А. Калмаков дал следующее определение информационному обществу: *«информационное общество – это глобальный экономико-политический, антропосоциальный и технологический проект, предполагающий управляемый переход к общественному устройству, при котором доминирующую роль во всех областях жизни будет играть система массовых коммуникаций (далее - СМК), реализованная с помощью компьютерных телекоммуникационных технологий, в частности, технологий интернета»*[1].

За внедрением новых технологий в производство, хранение и распространение информации СМК, пришло время новых средств массовой информации (далее – СМИ). Телевидение, радио и печать – это уже вчерашние традиционные СМИ, а новыми медиа принято считать мобильные платформы, социальные медиа, интернет-версии традиционных медиа. На фоне большого роста аудитории новых СМИ, традиционные СМИ теряют своих зрителей, слушателей и читателей. Так, по данным TNS за июнь 2016 года, социальные медиа «Вконтакте», Youtube, «Одноклассники» обогнали федеральный Первый канал по охвату аудитории[2].

Информация, размещаемая сегодня в современных средствах массовой информации все больше модифицируется в товар. Для того чтобы продать информацию, современные интернет-медиа борются за внимание зрителя. На медиаконференциях профессионалы из индустрии говорят о том, что необходимо руководствоваться законами драматургии при создании

мультимедийного контента в цифровой среде. Логика, верстка, расположение модулей, тип и размер картинки – все должно работать ради «якорей», чтобы читатель смог до конца прочесть материал и не переключиться. Как и в драматургии, размещаемый контент должен вызывать эмоциональный отклик у зрителя. А эмоции уже конвертируются в просмотры, лайки, комментарии и перепосты в социальных сетях, на количество которых обращают внимания рекламодатели – один из основных источников доходов современных медиа.

Отклик аудитории в виде огромного количества лайков, перепостов и комментариев используются крупными медиакомпаниями для исследования мнения аудитории. В то время, как показатели просмотров, лайков, дизлайков можно легко посчитать доступными инструментами, размещаемые в комментариях мнения читателей остаются неисследованными по причине неструктурированности текста. Анализ этой информации позволил бы медиакомпаниям повысить качество своего контента, выделить целевую аудиторию, определить настрой масс по отношению к контенту и компании в целом.

Автоматическое решение данной задачи в прошлом было невозможно. Сегодня же активное развитие компьютерной лингвистики позволяет извлекать информацию из текстов при помощи компьютерных технологий и математических моделей. Одним из напрапвлений данной дисциплины является задача определения эмоциональной окраски текста (анализ тональности текста, контент анализ, сентимент-анализ).

Анализ тональности текста – набор методов для определения эмоциональной окраски лексики текстов, эмоций автора по отношению к объекту и других свойств. Решение этой задачи компьютерной лингвистики позволит понимать текстовую информацию и упростит дальнейшее использование данных, полученных в результате ее систематизации и обработки.

Технология сентимент-анализа находит широкое применение у крупных компаний – владельцев брендов для анализа социальных медиа. Современные приложения сентимент-анализа дают возможность не только оценить тональность высказываний о бренде, но и получить целый ряд дополнительных инструментов, упрощающих управление социальной аудиторией, интересующейся брендом, установление контактов, обмен информацией, влияние на взращивание социального контента, поиск лидеров мнений социального сообщества, снабжение их информацией и привлечение к продвижению бренда.

Анализ тональности также применяется в области переводов текстов на другой язык, в котором первичная обработка текста повышает качество перевода. Методы анализа тональности могут применяться в разработке рекомендательной системы, которая будет советовать пользователю товары или услуги. Также стоит упомянуть, что технологии сентимент-анализа могут быть полезны политическим партиям, службам разведки. Применение таких технологий позволит им изучать мнения пользователей об определенном кандидате или событии.

Существуют несколько подходов для определения тональности текстов[3]:

- на основе правил;
- с помощью тональных словарей;
- машинное обучение с учителем;
- машинное обучение без учителя.

В первом подходе анализ текста проводится на основе заранее составленных набора правил. В рамках второго подхода каждому слову из текста присваивается тональность со значением тональности из тонального словаря (если оно присутствует в словаре). Общая тональность вычисляется как среднее арифметическое всех значений.

Третий подход обеспечивает высокую точность оценки текста. На

основе обучающей выборки классификатор самостоятельно выделяет признаки, влияющие на тональность. Таким образом, проблема зависимости от предметной области решается с помощью использования обучающей выборки из той же области. В четвертом подходе не требуется обучающая выборка для классификатора, но точность алгоритма ниже чем у алгоритмов, основанных на обучении с учителем.

Таким образом, поскольку основанные на обучении с учителем подходы показывают более высокие результаты при анализе текстов из социальных медиа и сайтов с рецензиями, мнениями, в данной работе будет подробно описаны именно эти методы и используемые в них способы представления данных. Кроме того, в работе будут использоваться некоторые приемы для улучшений точности работы алгоритмов.

Постановка задачи

Целью выпускной квалификационной работы является разработка метода автоматического определения тональности русскоязычных текстов-комментариев к видеозаписям на основе методов машинного обучения с учителем.

Для достижения поставленной цели необходимо решить следующие задачи:

1) формирование корпуса текстов-комментариев;
2) предварительная обработка данных - удаление нежелательных символов (шумов);
3) изучение нескольких векторных моделей представления текста (биграммы, униграммы) и функций весов (бинарная, частотная). Выбор оптимального сочетания модели языка и весов, которое даст наилучший показатель точности на данной выборке;
4) исследовать и реализовать несколько классифицирующих алгоритмов (в т.ч. наивный Байесовский алгоритм и метод опорных векторов).
5) проверка эффективности алгоритмов, выбор алгоритма, дающего наибольший уровень точности на конкретной обучающей выборке

Каждый рассматриваемый комментарий должен быть отнесен к одному из двух классов: положительная оценка или отрицательная оценка. Не должно возникать конфликтных ситуаций, при которых комментарий невозможно отнести ни к одному из двух классов. Точность разработанного метода классификации тональности текстов должна быть не менее 60% в пределах исследуемого корпуса.

Обзор литературы

Задача сентимент-анализа возникла относительно недавно и оптимальных подходов к ее решению на данный момент не существует. Все задачи, связанные с обработкой естественных языков являются сложными и неоднозначными. Несмотря на трудности, связанные с проблемой анализа тональности, имеется некоторое количество научных работ, посвященных данной тематике.

Начнем с того, что сентимент-анализ как наука начал развиваться с 2000-х годов. Термин «sentiment analysis» впервые был упомянут в работе [4], а выражение «анализ мнений» (с англ. «opinion mining») – в [5].

Котельников Е.В. и Клековкина М.В. в своем исследовании [6] представили методы автоматической обработки текстов на основе методов машинного обучения. Рассматривались несколько методов машинного обучения: метод опорных векторов, наивный Байесовский классификатор, метод ключевых слов и т.д. В ходе исследования были попытки найти оптимальный вариант векторной модели представления текстов и наилучший для этой модели классификатор. Экспериментально они определили, что использование бинарной модели с косинусной нормализацией без обучения и метода, комбинирующего использование ключевых слов и SVM, дает наилучшие результаты.

В исследовании Tang C. и соавт. [7] предлагаются два метода типа «обучение с учителем» (class association rules и наивный Байесовский классификатор) для выявления характеристик продукта или услуги и классификации этих характеристик с целью составления краткого резюме отзывов потребителей. Результаты их работы показывают, что каждый метод хорош по-своему, это зависит от конкретной метрики эффективности. Эффективность этих методов составляет более 70% (значение точности).

Для русского языка сентимент-анализ начал развиваться совсем недавно. Ежегодно проходит международная конференция по компьютерной

лингвистике «Диалог» [34], в программу которой включены доклады по данной теме.

В исследовании [8] используется метод на тональных словарях и лингвистических правилах. В работе исследуется извлечение терминов из русскоязычных текстов при помощи графовых моделей [9] экспериментально определило оптимальный размер окна при составлении графа русскоязычного текста. В статье [10] предлагается метод классификации со словарем, веса элементов которого определялись при помощи генетического алгоритма с точностью классификации более 65%. А точность около 70% была достигнута в работе [11] при рассмотрении метода на основе входящих в текст оценочных слов.

Как было указано выше, исследователи часто комбинируют подходы для достижения наилучших результатов. Например, научная работа Васильева В.Г., Давыдова С. и Худяковой М.В. [12] использует лингвистический подход, дополненный методами машинного обучения для коррекции отдельных правил классификации путем обучения.

Большое количество исследований говорят о том, что лингвистический подход в решении задачи тональности дает хорошие результаты. Алгоритмы, основанные на правилах, дают более точные результаты, так как работа этих методов тесно связана с семантикой слов, в отличие от методов машинного обучения, оперирующих статистикой и теорией вероятности. Но, как уже было упомянуто, лингвистический подход обладает рядом серьезных недостатков. Согласно Tang и соавт. [7]:

“Большинство существующих методов полагается на инструменты обработки естественного языка, чтобы разобрать и проанализировать комментарии к интернет-обзорам. Но все же они предлагают плохую точность, потому что комментарии в интернет-обзорах имеют тенденцию быть менее формальными, чем тексты комментариев к новостным статьям или статьям в журнале. Много предложений в текстах содержат

грамматические ошибки и неизвестные элементы, которые не существуют в словарях”

Мы в этой работе анализируем тексты из комментариев в социальных медиа, то применение лингвистического подхода невозможно. Как уже было сказано, пользователи интернета не могут дать для исследования грамматически правильные тексты. А подход на лингвистических правилах может дать точные результаты при работе с текстами из научных и журнальных статей или других грамматически верных текстов.

Кроме того, подход, основанный на правилах, привязан к конкретному языку. В связи с вышесказанным, с целью получения хороших результатов в настоящем исследовании рассматривается подход, основанный на методах машинного обучения с учителем.

Обзор существующих решений

Не смотря на актуальность и перспективность данной задачи, существует не так мало уже готовых разработок, которые работают на анализе тональности русскоязычных текстов. В почти во всех разработках задача анализа тональности является частью большой задачи лингвистического анализа текстов больших объемов. Рассмотрим некоторые из решений, которые поддерживают обработку русскоязычных текстов:

1. ***Texterra[13]*** – масштабируемое решение для анализа текстов с открытым API, основанное на использовании баз знаний, извлекаемых из Веб-ресурсов. Задача анализа тональности в сервисе решается с использованием методов машинного обучения. Поддерживаются русский, английский и корейский языки.

2. ***Eureka Engine[14]*** - это высокоскоростная система лингвистического анализа текстов модульного типа, позволяющая извлекать новые знания и факты из неструктурированных данных огромных объемов. Модуль SentiFinder автоматически определяет тональность текста по заданному текстовому объекту по трем видам тональности: позитивной, негативной и нейтральной. В модуле предусмотрено определение двух типов тональности:
   - относительно заданного пользователем объекта;
   - автоматически определенного системой объекта на основе совокупности знаний о нем.

   Используются методы машинного обучения. Сервис является продуктом коммерческого использования. В системе кроме модуля определения тональности, реализованы также определение языка сообщения (24 языка), определение тематики (автоклассификация), выделение именованных сущностей и имен собственных (NER), нормализация слов, разметка частей речи (морфоанализ).

3. ***SentiStrength[15]*** **–** программа оценки силы тональностей. Работает на основе словаря эмоционально окрашенной лексики, все слова в котором закодированы от -5 до -1 для слов, выражающих отрицательные эмоции и от 1 до 5 для слов, выражающих положительные эмоции. Согласно данным в [16] точность данная программа по точности классификации положительных комментариев показывает результат в 60,6%, по отрицательным – 72,8%.
4. ***RCO Fact Extractor[17]*** – система, разработанная компанией RCO. Данное решение использует методы, основанные на правилах. Данная си- стема учитывает синтаксическую структуру текста и взаимодействие различных типов слов
5. ***Twitter Sentiment[18]*** **–** один из популярных сервисов для анализа текстов, публикуемых пользователями в социальной сети Twitter. В этой социальной сети пользователи зачастую пишут свое мнение, отзыв о событии или о товаре. Сервис использует методы машинного обучения и имеется свое API. Работа данного сервиса построена на анализе последних 100 твитов (записей) пользователей о товаре или о событии. Достаточно ввести слово и сервис проанализирует тексты, построив графики соотношения положительных и отрицательных твитов.

Глава 1. Формирование коллекции комментариев для обучающей выборки

Рассмотрим площадку сайта Youtube, на котором каждую минуту выкладывается около 400 часов UGC-видеоконтента (User Generated Content) [19]. В связи с тем, что современные пользователи все реже используют телевидение как источник информации, многие телеканалы развивают на этом сайте свои каналы и загружают свой контент. В США на данной площадке начали практику потребления контента по платной подписке. Данная площадка является перспективной как альтернатива вчерашнему телевизору с эфирным вещанием. Президент Google в регионе EMEA Мэтт Бриттин, отмечает, что [19]:

«В 2016 году среднее время просмотра YouTube-видео на экранах телевизоров выросло более чем в два раза…В России количество просмотров на Youtube выросло на 50%».

Youtube является перспективным, с точки зрения развития, контент-провайдером с собственными алгоритмами продвижения и монетизации контента[20]. Кроме того, показатели потребления контента на данном сайте могут быть полезны для более глубокой оценки потребления контента. С его помощью можно оценить что важно для аудитории, что волнует современное общество.

1.1. Общие сведения о собранной базе с комментариями

Для анализа была собрана коллекция комментариев к видеозаписям канала «Бизнес секреты» в социальной сети *Youtube.com* [21]. Данный канал относится к тематике “Бизнес”. Ведущий канала проводит интервью с известными предпринимателями на тему бизнеса. Видеоролики выходят регулярно, с периодичностью 1 раз в месяц. Средний показатель просмотров за первую неделю составляет 100 000 просмотров. Аудиторией канала является активная молодежь от 16-18 до 35-40 лет - молодые предприниматели. Это дает основания предполагать, что качество

размещаемых комментариев будет выше, чем у каналов развлекательного жанра, где возраст аудитории может быть ниже.

Для работ было отобрано 4 видеозаписи, к которым имеется 3860 комментариев, 14457 лайков, 4337 дизлайков.

	комментарии	лайки	дизлайки	просмотры
1 видео	1 781	2 613	3 531	103 471
2 видео	1 089	6 718	237	154 328
3 видео	501	2 386	305	89 851
4 видео	489	2 740	264	111 559
ИТОГО:	3 860	14 457	4 337	459 209

*Таблица 1.1.*Схема исследуемой выборки

1.2. Обучающая выборка

Для использования методов обучения с учителем требуется обучающая выборка. Обычно обучающее множество составляется из примеров той области, в которой будет применяться классификатор. Но в текстах комментариев в социальных медиа используются слова из разных предметных областей. В представленной работе будут рассматриваться комментарии к видеозаписям на сайте Youtube в рамках одного канала.

Для формирования обучающей выборки была проведена ручная экспертиза комментариев с разметкой на положительные и отрицательные. В экспертизе участвовало 3 человека, перед которыми стояла задача не только разметить комментарии на положительные и отрицательные, но и удалить нейтральные комментарии, комментарии, не относящиеся к теме публикации и комментарии со спамом.

Классификация комментариев по тональности будет осуществляться на два полярных класса – позитивные и негативные эмоциональные оценки. Таким образом, после проведенной работы для обучающей выборки удалось получить из общего количества 3860 комментариев корпус из 3193 комментариев (1925 положительных и 1268 отрицательных).

Глава 2. Определение тональности методами машинного обучения

2.1. Задача классификации при обучении с учителем

Задача текстовой классификации определяется следующим образом[22].

Пусть существует текст комментария $d \in \mathbb{X}$, где $\mathbb{X}$ – векторное пространство комментариев, и фиксированный набор классов $\mathbb{C} = \{c_1, c_2, \dots, c_n\}$.

Необходимо одним из выбранных методов из обучающей выборки (множества комментариев с заранее известными классами) получить классифицирующую функцию (классификатор) $\Gamma(\mathbb{D}) = \gamma$, которая отображает комментарии в класс $\gamma: \mathbb{X} \to \mathbb{C}$. В нашей задаче множество $\mathbb{C}$ состоит из двух элементов (положительный и отрицательный тип комментариев).

2.1.1. Векторная модель языка

Точность будущих классификаторов зависит от векторной модели. Представим все комментарии из обучающей и тестовой выборок в виде n-грамм, где n-граммы – последовательности слов длины n. Если n=1, то такая модель называется «униграммой», если n равно двум – «биграммой», трём – «триграммой», и так далее.

Для примера возьмем комментарий «Пригласите нормальных мужиков. Галицкого хотим!», для которого юниграмма будет состоять из элементов {«Пригласите», «нормальных», «мужиков», «Галицкого», «хотим»}, а биграммы – {«Пригласите_нормальных», «нормальных_мужиков», «мужиков_Галицкого», «Галицкого_хотим»}.

Таким образом каждый комментарий будет представлен в виде вектора $d = (w_1, w_2, \dots, w_{|V|})$, где V – множество всех уникальных термов из обучающей выборки, w_i — вес i-го терма.

Существует несколько способов взвешивания [23]: бинарные и частотные функции. В исследовании [24] Панг Л. выявил эффективность

бинарных функций взвешивания. Таким образом, наличие термина в комментарии важнее, чем его частота. Бинарные векторы представляются как последовательность нулей и единиц. Вес термина будет равен 1, если взятый термин из словаря выборки встречается в тексте, иначе – 0.

2.2. Классификаторы

Самой важной частью работы является выбор признаков для решения задачи. В рамках работы были реализованы два алгоритма классификации – наивный Байесовский и метод опорных векторов.

Данные алгоритмы были выбраны исходя из следующих факторов:

- Наивный Байесовский (Naïve Bayes Classifier, NBC) – один из самых простых в реализации методов, но показывающий высокую вычислительную эффективность как при обучении, так и при классификации.
- Метод опорных векторов (Support Vector Machines, SVM) более сложный в реализации и вычислении. В данном алгоритме возможна многоклассовая классификация.

2.2.1. Наивный Байесовский классификатор (NBC)

В основе наивного Байесовского классификатора лежит теорема Байеса[25]:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (2.1)$$

где

- $\boldsymbol{P(c|d)}$ – апостериорная вероятность того, что комментарий d принадлежит классу c;
- $\boldsymbol{P(d|c)}$ – вероятность появления комментария d в классе c;
- $\boldsymbol{P}(\mathrm{c})$ – априорная вероятность класса c;
- $\boldsymbol{P(d)}$ – вероятность появления комментария d;

$$C_{MAP} = arg \max_{c \in C} P(c|d) = arg \max_{c \in C} P(d|c)P(c), \quad (2.2)$$

где C_{MAP} – максимальная апостериорная вероятность класса.

2.2.2. Метод опорных векторов (SVM)

Метод опорных векторов – невероятностный бинарный линейный классификатор. В основе данного метода лежит задача по построению (оптимальной) разделяющей гиперплоскости. Таким образом, выборку комментариев можно разделить, если на ней можно построить линейный пороговый классификатор[25]:

$$sign\left(\sum_{i=1}^{m} w_i * x^i - w_0\right) = sign(< w, x > -w_0) \quad (2.3)$$

где $x = (x^1, \dots, x^n)$ – признаковое описание объекта x;

вектор $w = (w^1, \dots, w^n) \in \mathbb{R}^n$;

скалярный порог $w_0 \in \mathbb{R}$ - являются параметрами алгоритма.

Соответственно, задача состоит в подборе таких значений вектора *w,* при которых функционал, определяющий число ошибок равен нулю:

$$\sum_{i=1}^{m} [y_i(< w, x > -w_0) \leq 0] = 0 \quad (2.4)$$

где $< w, x >= w_0$– разделяющая гиперплоскость.

2.3. Метрика эффективности

В качестве метрики эффективности работы алгоритмов были выбраны точность и полнота[26]. Точность – это доля текстов действительно относящихся данному классу по отношению к количеству текстов, которые причислены классификатором к этому классу. Полнота – отношение найденных классификатором текстов одного класса, к числу всех текстов этого класса в тестовой выборке.

Таким образом, оценка качества классификации текстов-комментариев будет осуществляться так, как это представлено в Таблице 2.1, где:

1. TP – истинно-положительная оценка;
2. TN – истинно-отрицательная оценка;
3. FP – ложно-положительная оценка;

4. FN – ложно-отрицательная оценка;

База текстов		Экспертная оценка	
		Положительная	Отрицательная
Оценка классификатора	Положительная	TP	FP
	Отрицательная	FN	TN

Таблице 2.1. Оценка качества р*азбиения документов на классы*

Соответственно, для положительных текстов-комментариев, точность и полнота будут определяться следующим образом:

$$\text{Точность} = \frac{TP}{TP + FP}, \tag{2.5}$$

$$\text{Полнота} = \frac{TP}{TP + FN}, \tag{2.6}$$

В качестве меры, которая одновременно учитывает и полноту, и точность, будет использоваться F-мера, которая определяется формулой (2.3)

$$F = 2 * \frac{\text{Точность} * \text{Полнота}}{\text{Точность} + \text{Полнота}}, \tag{2.7}$$

Глава 3. Программная реализация и анализ результатов

3.1. Язык программирования и среда разработки

В качестве языка программирования использовался Python версии 3.6.1. являющейся одним из распространенных инструментов для анализа данных. Выбор был сделан в пользу этого языка, поскольку в данном языке имеется библиотека Scikit-Learn, в которой реализовано большое количество алгоритмов машинного обучения[27]. Дополнительно были подключены следующие библиотеки:

- NumPy – библиотека для научных вычислений;
- Scipy – библиотека для научных и инженерных расчетов;
- Pymorphy2 – морфологический анализатор для работы с русским языком[28].

Эксперименты проводились на операционной системе MacOS Sierra с характеристиками, представленными в Таблице 3.1.

Характеристика системы	Значение
Производитель	Apple Inc.
Процессор	Intel(R) Core(TM) i5-4258U CPU @ 2.40GHz 2.40 GHz
ОЗУ	4 ГБ.
Тип системы	64-разрядная операционная система, процессор x64

Таблица 3.1. Характеристики системы

3.2. Формирование коллекций комментариев

Для формирования корпуса русскоязычных текстов-комментариев к видеозаписям были написаны два программых модуля: Youtube Downloader и Comment.Cutter. Youtube Downloader автоматически собирает странички с комментариями с сайта *Youtube* [35], а Comment.Cutter разбирает сохраненные с сайта html-страницы выбранных видеозаписей и получает из них

комментарии. В программах была использована библиотека Beatiful Soup[29] – инструмент для парсинга HTML/XML файлов.

При обработке комментариев им автоматически присваивается значение «1», предварительно все комментарии считаются положительными (Рисунок 3.1). Код программы Youtube Downloader приведен в Приложении 1, код Comment.Cutter – в Приложении 2.

Из собранных комментариев далее формировались обучающая и тестовая выборки. Перед работой программы автоматического определения тональности были также произведены операции по приведению всех текстов в нижний регистр и удалению всех лишних знаков. Для обучающей выборки для комментариев с положительными оценками оставляли указанную по умолчанию метку «1», а для отрицательных комментариев меняли метку на «0».

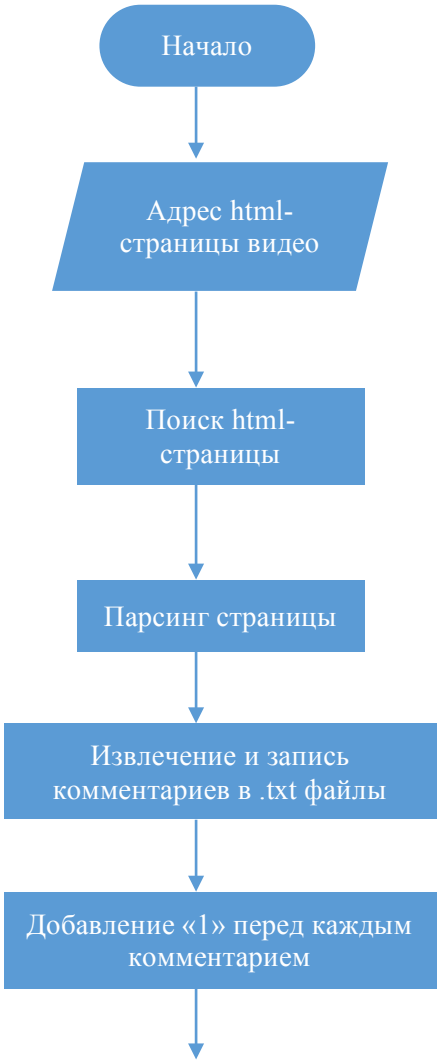

Рисунок 3.1. *Схема работы html-парсера*

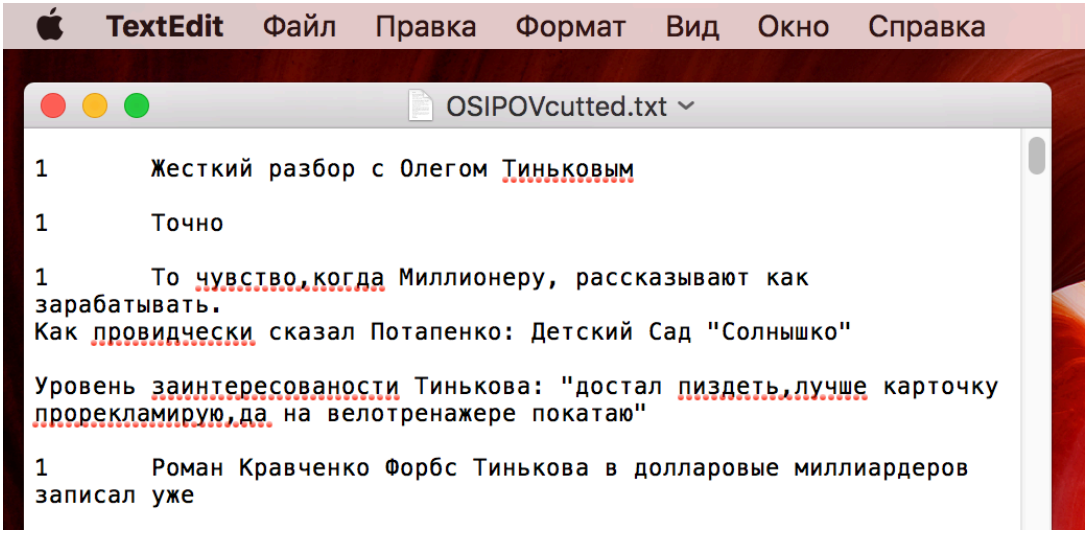

Рисунок 3.2. *Содержание текстового документа с комментариями.*

3.3. Результаты экспериментов

Все алгоритмы были реализованы на языке программирования Python с применением библиотеки Numpy для быстрых векторных операций. Для обучающей выборки было выбрано 85% из всего корпуса в 3193 комментариев. При использовании в обучающей выборке 50% корпуса, алгоритмы показали очень низкое качество работы. Тестирование проходило на всем корпусе размеченных комментариев. Для обучения и тестирования на каждой n-грамм на SVM тратилось время примерно 15 минут, на NBC – от 20 минут до 30 минут.

На результатах экспериментов видно, что среди выбранных классификаторов хорошие показатели точности достигаются при использовании униграммного представления данных в комбинации с наивным Байесовским классификатором. Данный метод показал точность равную 70%.

Алгоритм		Корпус		
Классификатор	n-грамм	Точность	Полнота	F-мера
SVM	1	66%	70%	65%

	2	64%	68%	60%
	1+2	62%	68%	57%
NBC	1	**70%**	**71%**	**70%**
	2	66%	69%	66%
	1+2	63%	69%	62

Таблица 3.2. Значения эффективности алгоритмов

В таблице 3.2 и 3.3 показаны значения эффективности методов определения тональности текста – методов опорных векторов и Наивного Байесовского классификаторов на каждом из типов комментариев.

Тип рецензии	Точность	Полнота	F-мера
Положительная	**78%**	83%	80%
Отрицательная	51%	42%	46%

Таблица 3.3. Характеристики эффективности Наивного-Байесовского классификатора

Тип рецензии	Точность	Полнота	F-мера
Положительная	72%	91%	80%
Отрицательная	**56%**	23%	32%

Таблица 3.4. Характеристики эффективности метода опорных векторов

Можно заметить, что положительные комментарии классифицируются точнее, чем отрицательные. Причиной этого могут быть слишком простые функции, использование для определения в комментариях иронии, сарказма, двусмысленных выражений; орфографические ошибки.

Полный текст написанных программ для решения задачи можно посмотреть по ссылке: https://github.com/BakytbekN/Youtubee

3.4. Дальнейшее направление исследований

В Таблице 3.3. показаны сравнения результатов тональности комментариев, полученных в ходе решения задачи, с имеющимися показателями лайков/дизлайков. Соотношение положительных комментариев к отрицательным комментариям значительно отличается от соотношения лайков и дизлайков. Таким образом, можно с уверенностью сказать, что показатели тональности могут быть отдельным измерением для анализа аудитории социальных медиа. А для контент-провайдера Youtube данный показатель может быть использован для продвижения контента пользователей на основе тональности комментариев.

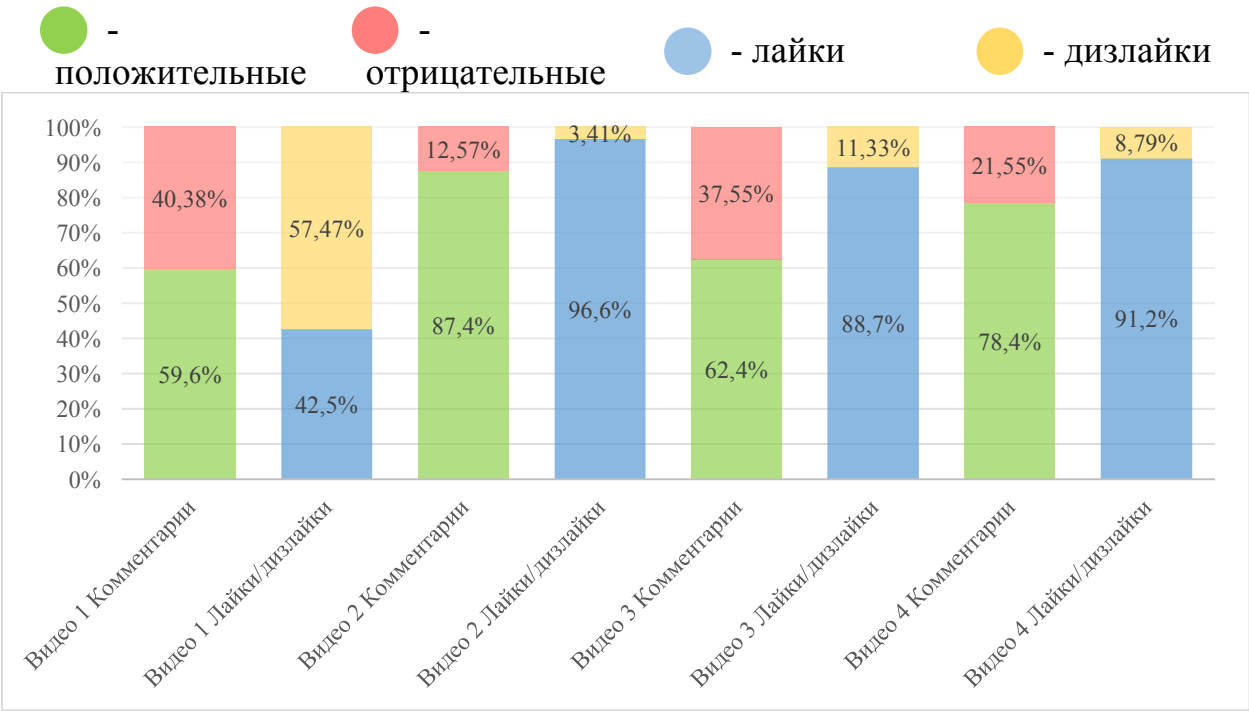

Таблица 3.3. Сравнение показателей тональности с лайками

Выводы

В представленной работе был реализован алгоритм анализа тональности русскоязычных текстов на основе методов машинного обучения.

Было замечено, что положительные комментарии классифицируются точнее, чем отрицательные. Все реализованные методы показали точность выше 60%. Невысокая точность данных подходов может быть связана со следующими причинами: слишком простая функция определения тональности предложений, использование в комментариях двусмысленных выражений, ироний; орфографические ошибки и личное отношение к теме эксперта при формировании корпуса для обучения и тестирования.

В рамках работы были решены следующие задачи:

1. Изучена проблема анализа тональности, проанализированы подходы для её решения.
2. Сформирован корпус русскоязычных комментариев к видеозаписям
3. Проведена нормализация данных: стемминг слов (приведение слов к их основам), удаление нежелательных символов, гиперссылок.
4. Реализованы алгоритмы классификации текстов по тональности – Наивный Байесовский классификатор и метод опорных векторов.
5. Проведено тестирование эффективности методов.

Заключение

Русскоязычные тексты отличаются от англоязычных сложной структурой. Реализованные методы машинного обучения при работе с англоязычными текстами показывают более хорошие результаты. Для получения более высоких результатов необходимы глубокие исследования.

Таким образом, можно выделить следующие направления для дальнейших исследований:

1) Дополнение алгоритма элементами лингвистического анализа;
2) Определение контекста обрабатываемого текста и сопоставление с темой видеозаписи, публикации;
3) Учет противительных союзов, лексем, хештегов.
4) Использование при сентимент-анализе эмодзи-смайлов, форм записей слов верхним регистром, как признаков выражения эмоциональности;
5) Распознавание спама и текстов, не относящихся к видеозаписи;

Список литературы

1. Калмыков А. А. Информационное общество // Экономико-математический энциклопедический словарь / под ред. В. И. Данилов-Данильян. М.: ИНФРА-М, 2003. С. 180-182.
2. Отчет TNS интернет аудитории// TNS Россия – Июнь, 2016. http://mediascope.net/services/media/media-audience/internet/information/ - Дата обращения: 24.05.2017
3. Обучаем компьютер чувствам (sentiment analysis по-русски) — http://habrahabr.ru/post/149605/ - Дата обращения: 24.05.2017
4. Nasukawa T., Yi J. Sentiment analysis: Capturing favorability using natural language processing // Proc. of the 2nd Int. Conf. on Knowledge capture (K-CAP), 2003. P. 7077.
5. Dave K., Lawrence St., Pennock D. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews // Proc. of the Int. Conf. on World Wide Web (WWW), 2003. P. 519528.
6. Котельников Е.В., Клековкина М.В. Автоматический анализ тональности текстов на основе методов машинного обучения. РОМИП 2011.
7. Tang, X. Yang, C., Wong, Y., Wei C. Understanding Online Consumer Review Opinions with Sentiment Analysis using Machine Learning // Pacific Asia Journal of the Association for Information Systems. – 2010. - № 3(2). – С. 73-89.
8. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Международная конференция по компьютерной лингвистике Диалог, 2011. С. 510 – 522.
9. Усталов Д. Извлечение терминов из русскоязычных текстов при помощи графовых моделей // Теория графов и приложений, 2012. С. 62-69.
10. Котельников Е.В., Клековкина М.В. Определение весов оценочных слов на основе генетического алгоритма в задаче анализа тональности

текстов // Программные продукты и системы, 2013. Вып. 4. С. 296–301.

11. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL) / Переславль-Залесский: изд-во «Университет города Переславль», 2012. С. 118–123.
12. Худякова М.В., Давыдов С., Васильев В.Г. Классификация отзывов пользователей с использованием фрагментных правил. РОМИП 2011.
13. ИСП РАН[Электронный ресурс]: TEXTERRA. Технология автоматического построения онтологий и семантического анализа текста. http://www.ispras.ru/technologies/texterra/ - Дата обращения: 24.05.2017
14. NL Pub[Электронный ресурс]: Eureka Engine. Режим доступа: https://nlpub.ru/Eureka_Engine - Дата обращения: 24.05.2017
15. SentiStrength [Электронный ресурс]: SentiStrength – sentiment strength detection in short texts. – Режим до- ступа: http://sentistrength.wlv.ac.uk/#About - Дата обращения: 24.05.2017
16. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., &Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558
17. RCO Fact Extractor SDK [Электронный ресурс]: RCO. http://www.rco.ru/product.asp?ob_ no=5047 - Дата обращения: 24.05.2017
18. Веб-сервис Twitter Sentiment [Электронный ресурс]: Режим доступа: http://www.sentiment140.com - Дата обращения: 24.05.2017
19. Тренды Youtube: Влогерский бум и коллаборация с ТВ // Adindex Print Edition. – 2017. – №28. https://adindex.ru/publication/analitics/100380/2017/04/12/159181.phtml Дата обращения: 24.05.2017
20. Продвижение видео на Youtube. SEO, аналитика, схемы и нюансы. – Habrahabr, 2016

https://habrahabr.ru/company/wargaming/blog/293070/
Дата обращения: 24.05.2017

21. Канал «Бизнес секреты» на Youtube https://www.youtube.com/user/BiZSekrety - Дата обращения: 24.05.2017
22. Manning D., Raghavan P., Sch¨utze H. Introduction to Information Retrieval // Cambridge University Press, 2008
23. Википедия. Векторная модель — http://ru.wikipedia.org/wiki/Векторная_модель – Дата обращения: 24.05.2017
24. Pang L. Thumbs up Sentiment Classification using Machine Learning Techniques // Proceedings of EMNLP (2002).
25. К.В. Воронцов. Математические методы обучения по прецедентам (теория обучения машин), 2011.
26. Статья о характеристиках измерения эффективности классификатора https://en.wikipedia.org/wiki/Precision_and_recall - Дата обращения: 24.05.2017
27. Fabian P., Gaël V., Alexandre G., Vincent M., Bertrand T., Olivier G., Mathieu B., Peter P., Ron W., Vincent D., Jake V., Alexandre P., David C., Matthieu B., Matthieu P., Édouard D.. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research 12 (2011) 2825-2830
28. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332 (2015).
29. Документация Beatiful Soup 3 https://www.crummy.com/software/BeautifulSoup/bs3/documentation.html – Дата обращения: 24.05.2017
30. Сарбасова А. Н. Исследование методов сентимент-анализа русскоязычных текстов // Молодой ученый. — 2015. — №8. — С. 143-146.

31. Прохоров А., Керимов А., Сентимент-анализ и продвижение в социальных медиа // Компьютер Пресс. – 2012. Июль http://compress.ru/article.aspx?id=23115 – Дата обращения: 24.05.2017
32. Котельников Е.В., Пескишева Т.А., Пестов О.А. Параллельный выбор параметров классификатора для анализа тональности текстов // Вопросы современной науки и практики. Университет им. В.И. Вернадского, 2012. Вып. 4(39). С. 100-106.
33. Черных А. Мир современных медиа.— М.: Издательский дом «Территория будущего», 2007. (Серия «Университетская библиотека Александра Пого- рельского»).—312 с.
34. Международная конференция по компьютерной лингвистике «Диалог» http://www.dialog-21.ru/
35. Youtube[электронный ресурс] – контент-провайдер. http://www.youtube.com – Дата обращения: 24.05.2017

Приложение 1. Код Youtube Downloader

```
impor
t time
        from selenium import webdriver
        from selenium.webdriver.common.by import By
        from selenium.webdriver.support.ui import WebDriverWait
        from selenium.webdriver.support import expected_conditions as EC
        from selenium.common.exceptions import TimeoutException
        from bs4 import BeautifulSoup

        def check_exists_by_xpath(xpath,driver):
            try:
                driver.find_element_by_xpath(xpath)
            except NoSuchElementException:
                return False
            return True

        def init_driver():
            driver = webdriver.Firefox()
            driver.wait = WebDriverWait(driver, 5)
            return driver


        def lookup(driver,executeScript):

        driver.get('https://www.youtube.com/watch?v=_ndPr4k9ZSQ&t=243s'
        )
            try:
                while(True):
                    time.sleep(3)# google podumaet bot i zabanit tvoj IP... mozhet
        byt' takoe TOCHNO NE UVEREN
                    button =
        driver.wait.until(EC.presence_of_element_located(((By.XPATH,
        u"//div[@id='comment-section-renderer']/button"))))
                    if button.is_enabled():
                        button.click()
                    currentSiteName=driver.title
            except TimeoutException:
                print("should print now")
                driver.execute_script(executeScript)# Commenty na comment ne
        zagruzhayutsya vmeste nuzhno podgruzhat potom jetot javascript
        vyzovet potom i podgruzit
                time.sleep(40)
```

```
        content= driver.page_source
        soup=BeautifulSoup(content, 'html.parser')
        html = soup.prettify("utf-8")
        print("prittified")
        with open("output1.html", "wb") as file:
            file.write(html)
        print("fileCreated")


if __name__ == "__main__":
    javascriptExecutable=";"
    with open("JavaScriptCommentLoader.txt") as f:
        for line in f:
            javascriptExecutable+=line
    driver = init_driver()
    lookup(driver,javascriptExecutable)
    time.sleep(5)
    driver.quit()
```

Приложение 2. Код comment-cutter`a

```
from bs4
import
BeautifulSoup
                import codecs
                #
                names=["DENIS","FILEV","OSIPOV","VADIM"]


                def cutter(name):
                        tekst=""
                        f = codecs.open(name+".txt", 'r', 'UTF-8')
                        for line in f:
                                tekst+=line
                        f.close()
                        print(len(tekst))
                        vernut=""
                        # with open("DENIS.html", 'rb') as f:
                                # lines = [x.decode('utf8').strip() for x in
                f.readlines()]
                        # for i in lines:
                                # tekst += str(i) + "-"
                        soup = BeautifulSoup(tekst, 'html.parser')
                        element=soup.findAll("div", { "class" : "comment-
                renderer-text-content" })
                        for hit in soup.findAll("div", { "class" : "comment-
                renderer-text-content" }):
                                vernut+="1\t"+hit.get_text().lstrip()+"\r\n"
                        #    vernut+=hit.get_text().lstrip()+"\r\n"
                        with open(name+"cutted.txt", "wb") as fileF:
                                fileF.write(vernut.encode("utf-8"))


                for name in names:
                        cutter(name)
```